DOCUMENT RESUME

ED 300 449 TM 012 498

AUTHOR Green, Donald Ross; And Others
TITLE Experiences in the Application of Item Response Theory in Test Construction.
PUB DATE Apr 88
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Computer Assisted Testing; Difficulty Level; Elementary Secondary Education; Error of Measurement; *Latent Trait Theory; Models; Test Bias; *Test Construction
IDENTIFIERS *Three Parameter Model

ABSTRACT

Potential benefits of using item response theory in test construction are evaluated, based on the experience and evidence accumulated during 9 years of using a three-parameter model in the construction of major achievement batteries. Specific benefits covered include obtaining sample-free item calibrations and item-free person measurement, automatically equating various subsets of a pool of calibrated items, decreasing standard errors of scores without increasing the number of items used, assessing item bias independently of item difficulty, determining the adequacy of the tryout pool of items, establishing computer-generated ideal tests drawn from item pools as targets for test developers, and controlling the standard error of the test selected at any desired set of score levels. The matter of test dimensionality is also addressed. Although the three-parameter model appears to be very powerful, users must be cautious in its application due to the fact that the assumptions of the model are rarely fully met. Six graphs conclude the document. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300449

Experiences in the Application of Item Response Theory in Test Construction

by

Donald Ross Green

Wendy M. Yen

George R. Burket

CTB/McGraw-Hill

Paper Presented at the Annual Meeting of the National Council of Measurement in Education

New Orleans, April 6, 1988

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Donald Ross Green

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM012498

Certain potential benefits of using item response theory in test construction are well known and commonly cited in textbooks and articles (e.g., Allen & Yen, 1979; Crocker & Algina, 1987; Lord, 1980; Yen, 1983). They include the possibility of:

(a) getting so-called "sample free" item calibrations, which has been interpreted to mean that the need for representative samples in item tryouts is essentially eliminated;

(b) so-called "item-free" person measurement, which has been interpreted to mean that any subset of a pool of calibrated items can produce trait estimates as valid as the estimates from any other subset of the pool;

(c) automatically equating of these various subsets; and

(d) decreasing the standard errors of scores without increasing the number of items used by using item pattern scoring.

(e) Assessing item bias (or differential item functioning) independently of difficulty in a manner consistent with item selection.

Other benefits not so commonly cited are the possibility of:

(a) being able to determine just how adequate the tryout pool of items may be;

(b) setting up computer generated "ideal" tests drawn from item pools as target for test developers; and

(c) controlling the standard error of the test selected at any desired set of score levels.

In this paper these potential benefits are evaluated using the experience and evidence accumulated during nine years of using a three parameter model in the construction of major achievement batteries. A number of cautions and limitations in realizing these benefits are discussed and a number of issues in need of further research are mentioned.

IRT versus Traditional Test Construction Procedures

It is well to realize that the basic procedures of test construction are altered little if at all by the use of item response theory. A quick rundown of the major steps in constructing an achievement battery may make this more apparent.

The procedures for developing achievement tests may be summarized as follows:

1. Specifications: Content specifications are established. IRT does not influence this basic step.

2. Item Writing: Items are written (or found). IRT does not influence this step.

3. Item Tryouts: Items are tried out. IRT does not influence this step.

4. Item Selection: A set of items tried out are selected and assembled to form the test. The analyses done for this step are different than traditional analyses but the underlying ideas are the same, i.e., the editor selects the statistically best items within the constraints of content validity and the desired measurement characteristics of the test. In both approaches item difficulty and item discrimination are major selection criteria.

5. Editing: The standardization edition is produced and edited for content, format, etc. (Note: major editing efforts, including external content and bias reviews, occur during and after Step 2). IRT does not influence editorial procedures.

6. Standardization and Norming: The samples and testing conditions are not influenced by IRT. The scaling procedures used with the standardization data are different but the norms themselves are not. Norms for traditional and IRT based tests have the same kinds of derived scores (i.e., percentile ranks, NCEs, GEs, scale scores) and these scores have the same meaning.

7. Technical Studies: The equating and other studies -- especially vertical equating studies -- are done in different ways with IRT but the same kinds of studies are needed with both approaches.

8. Final Edition: Again IRT does not change matters; the final edition should be the same as what was standardized. Occasionally one finds it necessary to delete a defective item, but, as do traditional procedures, IRT constrains the possibility of altering or further editing the items used.

As this list shows test construction based on item response theory procedures and traditional procedures are basically the same. The benefits of using item response discussed below should be understood in that context. Each of the potential benefits mentioned in the beginning will be considered separately.

Sample-Free Item Calibrations

Regardless of the approach to test construction being used it usually has been accepted that it is not necessary to obtain truly representative samples for item tryouts because the data are used to identify the better items rather than provide final item statistics or scores. At CTB we have always used grab

samples rather than probability samples for tryouts (in contrast to norming) but have required that the classes be heterogeneous in nature and that a variety of school systems in various parts of the country be included.

Item response theory does not change these requirements. Some items can have very different parameters in particular schools and school districts because of specific curricular and instructional patterns (Yen, Green, & Burket, 1987). For example, a test of Arithmetic Computation may appear to be essentially unidimensional among groups taught much the same things, but given, say, some third grade students who have been introduced to multiplication and others who have not, the parameter estimates the multiplication items would get from the students that had been taught multiplication would differ sharply from the estimates derived from the group that had not. Thus if most districts do teach multiplication in the third grade it is necessary to use such districts in the tryouts to get valid indices of how such items will function. The procedure, therefore, is not fully sample free. Of course traditional item difficulties are affected the same way.

It may be noted that ability estimates are more robust than item parameters (Yen, 1980), but in item tryouts, item parameters are the focus of interest.

One still needs to include the full expected range of abilities in the tryout sample, even if the test is undimensional, in order to make it possible to estimate the parameters of very easy and very hard items.

It is also still useful to use tests from prior editions as anchors to get both estimates of national p-values for the items being tried out and estimates of the range of locations needed for the various levels of the tests being constructed (CTB/McGraw-Hill, 1987).

Nevertheless, we believe that building tests by calibrating items from tryout data using a three parameter model is a more effective approach than relying on traditional item statistics alone (Yen, 1983). The IRT procedure does allow one to have a better understanding during construction of the consequences of selecting particular items - partly because there is some degree of independence of the data from the particulars of the sample, in spite of the limitations just cited. Given adequate sample size, the estimates of item parameters obtained are less sensitive to sample variations than are traditional item statistics such as p-values, point biserials, and the like.

More to the point we have found that the traditional and IRT approaches tend to compliment each other. For example, sometimes the point biserials for all answer choices across several adjacent grade groups can help one understand the item characteristic curve derived from the pooled data and thus provide a sounder basis for accepting or rejecting the item. On the other hand

traditional item statistics based on a sample that is either atypically able or inept will suggest that certain items near the top or bottom of the ability range of the sample are poor items; the parameters of these items may well contradict this conclusion.

Item-Free Person Measurement

As with sample-free item calibration, content variations can have significant effects on ability estimates if these variations are reflected in school curricula and instruction (Yen, Green, & Burket, 1987). The situation previously described concerning a test of Arithmetic Computation permits a parallel statement about ability estimates. The ability estimates these students would get from a set of items that included multiplication would differ sharply from the estimates derived from an otherwise identical item set that did not. Note that before the first group was introduced to multiplication both tests might well produce reasonably equivalent trait estimates. Again the situation is the same for traditional procedures.

As noted previously the effects of content variations on trait estimates tend to be substantially smaller than the effects of content variations on item parameters (Yen, 1980).

It appears safe to say that item sets differing in difficulty and/or discrimination but not in content coverage will produce equivalent trait estimates although perhaps having different standard errors of measurement. It must be noted, however, that it is not entirely clear when content really differs. This is an area where research is needed.

Test Equating

The preceding statement is one way of saying that different sets of items calibrated to the same scale can "automatically" be considered to be equated, i.e., that they will produce tau-equivalent scores. While all the cautions about content equivalence discussed in the previous sections need to be observed before considering the scores from two such tests to be equivalent, it may be apparent why equating procedures based on item response theory are now widely used.

We have found such procedures convenient and effective for both interlevel and interform equating (CTB/McGraw-Hill, 1987). Clearly interlevel equating is the most difficult challenge in equating test developers face. Yen (1985) has shown that an IRT procedure produces results as good as those obtained with a Thurstonian equipercentile procedure. It is suspected that procedures other than these two are inferior, but there is clearly room for research on this issue.

Score Reliability

Item pattern scoring which entails finding maximum likelihood trait estimates from item response vectors (Lord, 1980, pp. 58-59) offers the possibility of reduced standard errors of measurement. Yen (1984) has shown that a maximum likelihood scoring procedure based on a number correct score produces trait estimates that are tau-equivalent to the pattern procedure. She also demonstrated, however, that the number correct procedure yields higher standard errors of measurement especially for scores at the lower end of the scale.

Arguments have been made verbally (but to my knowledge not published) that pattern scoring, which takes into account guessing as well as item difficulty and item discrimination, will handicap low scoring minority students. Green and Yen (1983) demonstrated that this result does not occur. Specifically they found that pattern scoring did not produce lower average scores for blacks at any grade level or in any subject than they obtained when number correct scores were used.

Another aspect of scoring that the use of item response theory has permitted us to improve is the reporting of scores on the specific objectives included in each of the tests. It has been common practice for more than thirty years (e.g., see CAT 1957 Analysis of Learning Difficulties) to try to get more detailed information from standardized tests by looking at the results for individual items or the small groups of items which refer to relatively specific objectives. Clearly this approach not only tends to deny the unidimensionality of the test but runs afoul of reliability standards, i.e., single items or small sets of items necessarily produce scores with very large errors of measurement. The move to report item and/or objective scores began with reports showing results item by item for individual students (some published tests still offer this "feature" in their score reports). Then came the instructional objective, still usually too short to have any satisfactory degree of reliability.

In 1977, Burket developed a Bayesian procedure for objective scores which utilized the score from the whole test as a prior and permitted sharply reduced standard errors of measurement (CTB/McGraw-Hill, 1979). As initially implemented this procedure tended to make the test score the major or even the sole determiner of the estimated objective score. Subsequent modifications of this Bayesian approach using item response theory have made it possible to report objective scores in a practical, useful and adequately reliable manner. The procedure has been described by Yen (1987).

Assessing Item Bias

Attempts to assess item bias or differential item functioning using traditional item statistics frequently encounter problems because of the differential level of performances among the groups being considered. The use of procedures based on item response theory can reduce the impact of this problem. Consequently item response theory approaches to dealing with item bias have begun to be widely used. There are reasons to argue that three parameter approaches are potentially very effective procedures (Ironson, 1980). Most of these approaches, however, entail assessing the parameters for each group of interest and therefore require larger samples than often is feasible. One solution to dealing with this parameter has been suggested by Linn and Harnish (1981).

We have used this approach and the results appear to make sense; for example the numbers of "biased" items so identified are always much smaller for random subsets of the overall group than for self identified ethnic subgroups. Nevertheless, more information about just how large the groups need to be get stable results for items would be helpful as would information about the meaning of results when the expected and obtained curves cross.

Other Benefits

Finally there are the set of "other benefits" referred to at the beginning of this paper which have been realized at CTB through the use of computer programs. The latest and most useful of these is a system of programs designed to be used by editors for item selection via microcomputers developed by Burket. Since the benefits referred to all are realized via this program they will not be considered separately. Only the features of the programs that make use of item response theory will be described.

The first program enables the editor to assemble a pool of all eligible items eliminating any that do not meet predetermined criteria with respect to location and discrimination as well as content.

The second program is used to do the actual item selection. The characteristics of the pool of items can be displayed graphically as illustrated in Figure 1. The baseline represents the trait scale and the top and bottom of the "nominal range", i.e. the predetermined range of locations within which the test being selected is intended to function. The distribution of item locations is displayed just below the baseline (there are about 200 items in the pool and forty in the test). The lower curve indicates the standard errors measurement at each point on the scale if the entire pool is used. The upper curve represents the forty item test chosen from the pool. The lower the curve the greater the amount of information being provided; thus lower is better.

The curve in Figure 1 is for a very good pool. Figures 2, 3, and 4 illustrate various problems such as gaps in the locations of pool items, inadequacies at the top or bottom of the range, and irregular curves caused by very poor items around some location. Since the pool contains all the information one can possibly obtain in any test selected from it, the pool curve can be compared to that for any for any test selected indicating where improvements are likely to be possible.

This item selection program can automatically select "optimal" tests from the pool within various constraints. Optimal may be defined in any of several ways:

In one procedure the nominal range is divided into as many points as there are items to be selected. At each point that item is selected which provides the maximum information at that point. The order in which the points are taken affects the outcome and so alternatives are offered.

A second strategy, which sometimes gives better results, selects the item that contributes the maximum information at the top or bottom of the nominal range or at an interior point which is a local minimum. Arbitrary rules are used for alternating between these three points. If good measurement near the floor and ceiling of the test is required the outside points can be given some priority.

A third strategy divides the nominal range into six equal intervals and then simply selects the items with the highest discriminations in these intervals aiming for an equal number of items in each interval.

Each of the approaches yields a somewhat different "optimal" test. The standard error curves and the distribution of locations of the items chosen can be displayed and compared with each other and with the pool. Figure 5 shows the curves for a pair of tests automatically selected from the pool in Figure 1. It should noted that each selection appears in a different color making comparisons easy.

Constraints and requirements in terms of content, e.g. objective representation, reading passages or other displays that may be used, and the like, can be imposed on this automatic selection process. Nevertheless, even in simple situations it is unlikely that the result will meet the full approval of an editor. These program selections, however, are an efficient starting point for the work of the editor.

To aid the editor in making further choices the program can be asked to display the locations of all items color coded by objective. Similarly the item discriminations can be displayed as the amount of information each item offers at any specified location. This latter feature makes it simple to maximize reliability at any predetermined location.

The third program displays the results of a selection giving not only the standard error curve and the test characteristic function graphically but also numerically. Other summary pieces of information are given such as the number of poorly fitting items used, the average bias rating of the test and that of the pool. A useful feature in building multilevel tests is that one can display the standard error curves or test characteristic curves of successive levels of the test thus ensuring proper measurement all along the scale. Figure 6 shows three successive levels of a test indicating their overlap and unique contribution not just nominally but in terms of the amount of information provided.

In using this system in the last six months for a number of tests we are impressed with the way it enables us to capitalize on the strengths of an item pool and to build test rapidly. It does not of course find a way around a poor initial pool of items and it does not eliminate the need for strong editorial control of the final test. Nevertheless, we believe that it gives us very good control of the construction because of its basis in item response theory.

Multidimensionality

Finally, we would like to comment on the matter of test dimensionality. As is well known, item response theory as currently applied uses propositions which assume that the trait being measured is unidimensional. Almost as well known is that this assumption is not fully met in most, if not all, achievement tests. Some people hold that this fact precludes the effective use of item response theory (e.g., Traub, 1983). We believe our experience shows that not to be the case, although as we have already noted lack of unidimensionalty does put limitations on some of the potential benefits we have listed.

It should be recognized that classical approaches also assume the existence of a unidimensional trait but offer no way of detecting whether or not that assumption has been violated or the consequences when it has. In fact, one of the major differences between using traditional development procedures and item response theory procedures, is that when one makes use of item response theory one is forced to face the flaws in the test one is building because they are much more visible than is the case when one is using traditional procedures. While not everybody considers this a positive outcome because it may be unpleasant news, it does permit one to try to do better the next time around.

Conclusions

In sum we believe that our experiences in the course of test construction show that it is advantageous to use a three parameter item response model. The major benefits are the greater

knowledge of, and therefore control of, the characteristics of the test. Nevertheless, it should be clear that the millennium has not arrived and one may produce an inferior test using item response theory almost as readily as one can produce an inferior test following traditional procedures; after all the basics are not really different. What is different is that, if one uses item response theory (and if one cares), one can become aware of some of the limitations of the test even before standardization, as well as before the validation studies that use the completed test.

We have noted that there are limitations and difficulties in using the model. The data requirements are substantial and if adequate data are lacking one might just as well follow traditional procedures since it will be immediately evident that attempts to apply the model will not work. Of course the traditional procedures do not compensate for inadequate data, but since they are relatively insensitive to such inadequacies they do not break down and so one can proceed, albeit blindly.

The assumptions of the model are rarely fully met and that fact calls for caution in its application. Ignoring the consequences that can arise from using the model when the assumptions are not fully met can have serious negative consequences as we have noted here - and explained elsewhere (Yen, Green, & Burket, 1987). We would like to repeat, however, that all tests that report a single score derived by summing correct responses in some way assume unidimensionality. The fact that traditional procedures do not make it possible to detect the consequences of violating this assumption does not make traditional approaches superior.

It is our fond hope that still better procedures are just around the corner.

References

Allen, Mary J. & Yen, Wendy M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company.

Crocker, Linda & Algina, James. (1986). Introduction to classical & modern test theory. New York: CBS College Publishing.

CTB/McGraw-Hill. (1979). California Achievement Tests Forms C and D technical report. Monterey, CA: CTB/McGraw-Hill.

CTB/McGraw-Hill. (1987). California Achievement Tests Forms E and F technical report. Monterey, CA: CTB/McGraw-Hill.

Green, Donald Ross & Yen, Wendy M. (1983, April). Number correct versus pattern scoring: Results for ethnic groups. Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Canada.

Ironson, Gail H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In Ronald A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 117-160). Baltimore, MD: The Johns Hopkins University Press.

Linn, R. L. & Harnish, D. L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Lord, Frederic M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Traub, Ross E. (1983). A priori considerations in choosing an item response model. In Ronald K. Hambleton (Ed.), Application of item response theory (pp. 57-70). Vancouver, British Columbia: Educational Research Institute of British Columbia.

Yen, Wendy M. (1980, Winter). The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17(4), 297-311.

Yen, Wendy M. (1983). Use of the three-parameter model in the development of a standardized achievement test. In Ronald K. Hambleton (Ed.), Application of item response theory (pp. 123-141). Vancouver, British Columbia: Educational Research Institute of British Columbia.

Yen, Wendy M. (1984, Summer). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic-model. Journal of Educational Measurement, 21(2), 93-111.

Yen, Wendy M. (1985, April). Tau quivalence of vertical equatings using three-parameter item response theory and Thurstonian procedures. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Yen, Wendy M. (1987, June). A Bayesian/IRT Index of Objective Performance. Paper presented at the meeting of the Psychometric Society, Montreal, Canada.

Yen, Wendy M., Green, Donald Ross, & Burket, George R. (1987, Spring). Valid normative information from customized achievement tests. Educational Measurement: Issues & Practice, 6(1), 7-13.

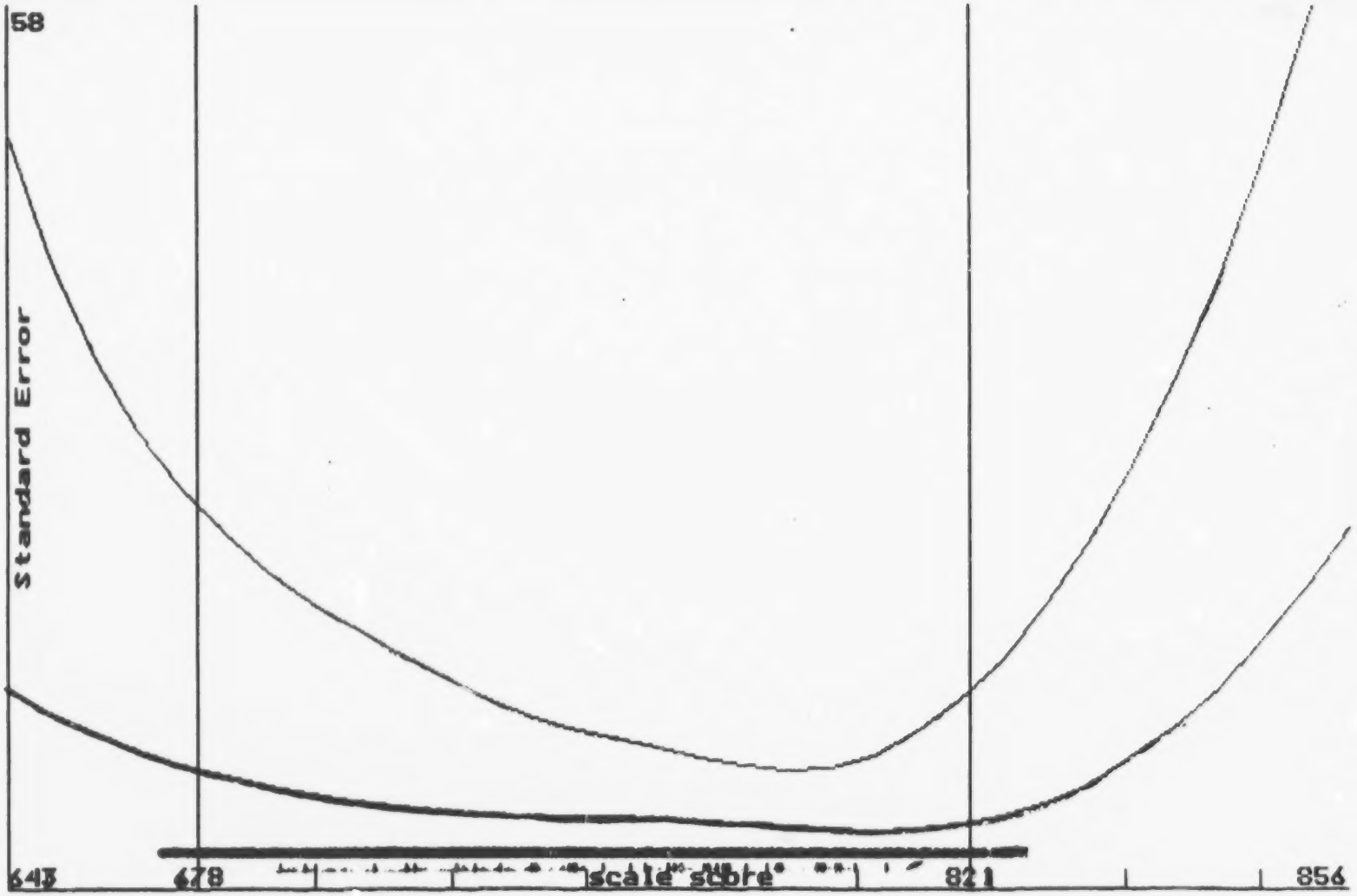

Figure 1. Standard Error Curves for an Item Pool and the Selected Test

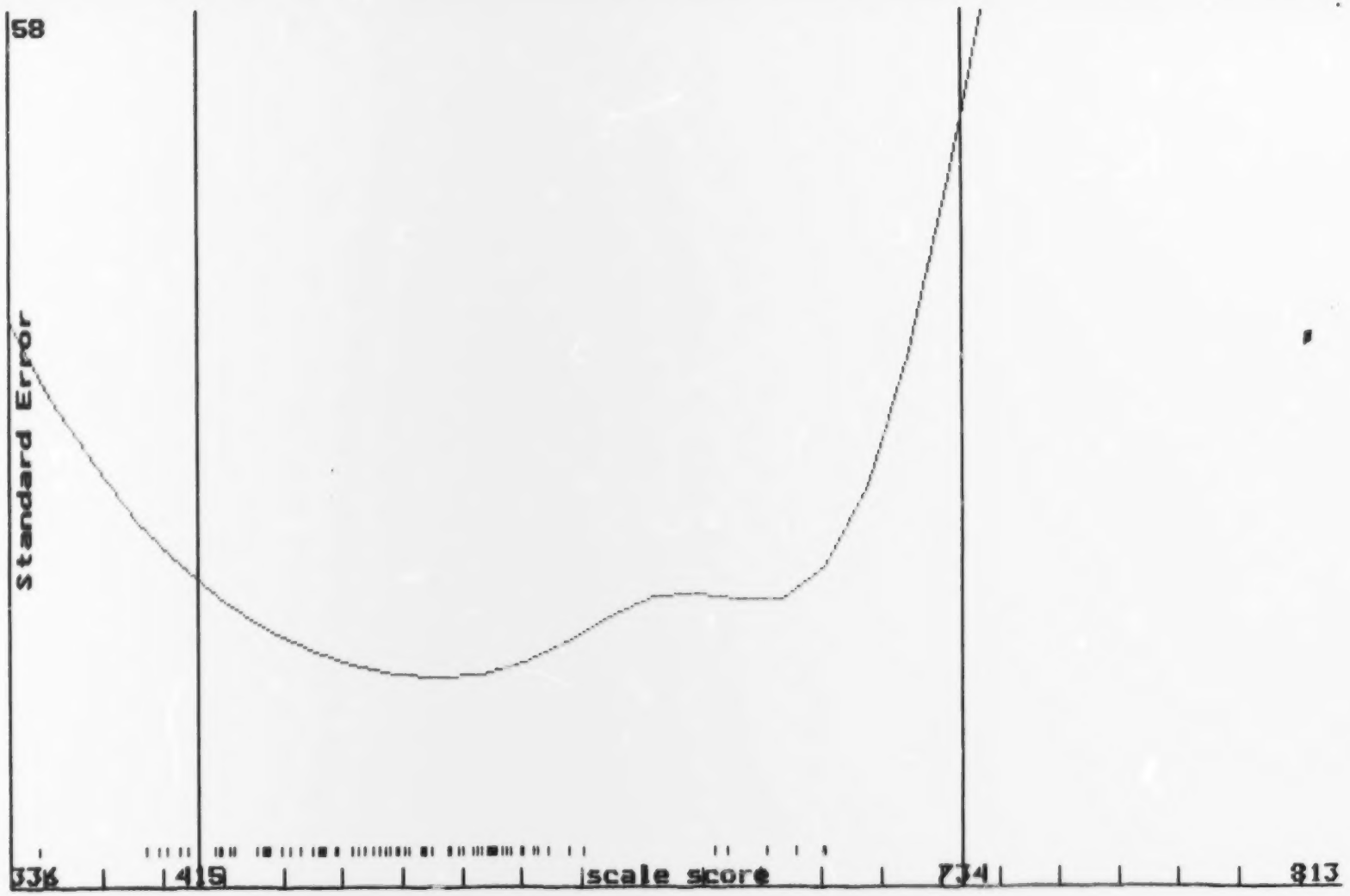

Figure 2. Standard Error Curve for an Item Pool Lacking Difficult Items

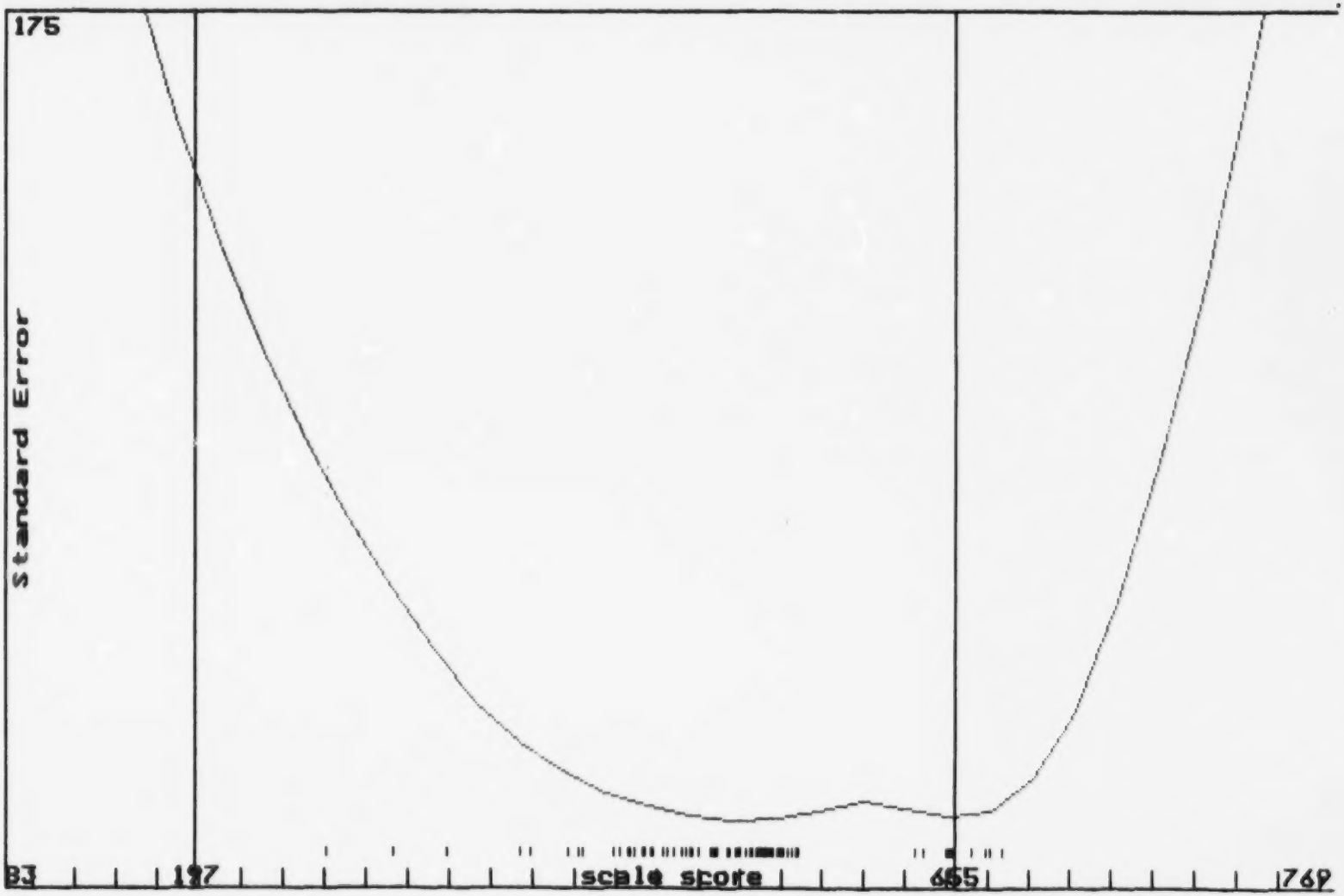

Figure 3. Standard Error Curve for an Item Pool Lacking Easy Items

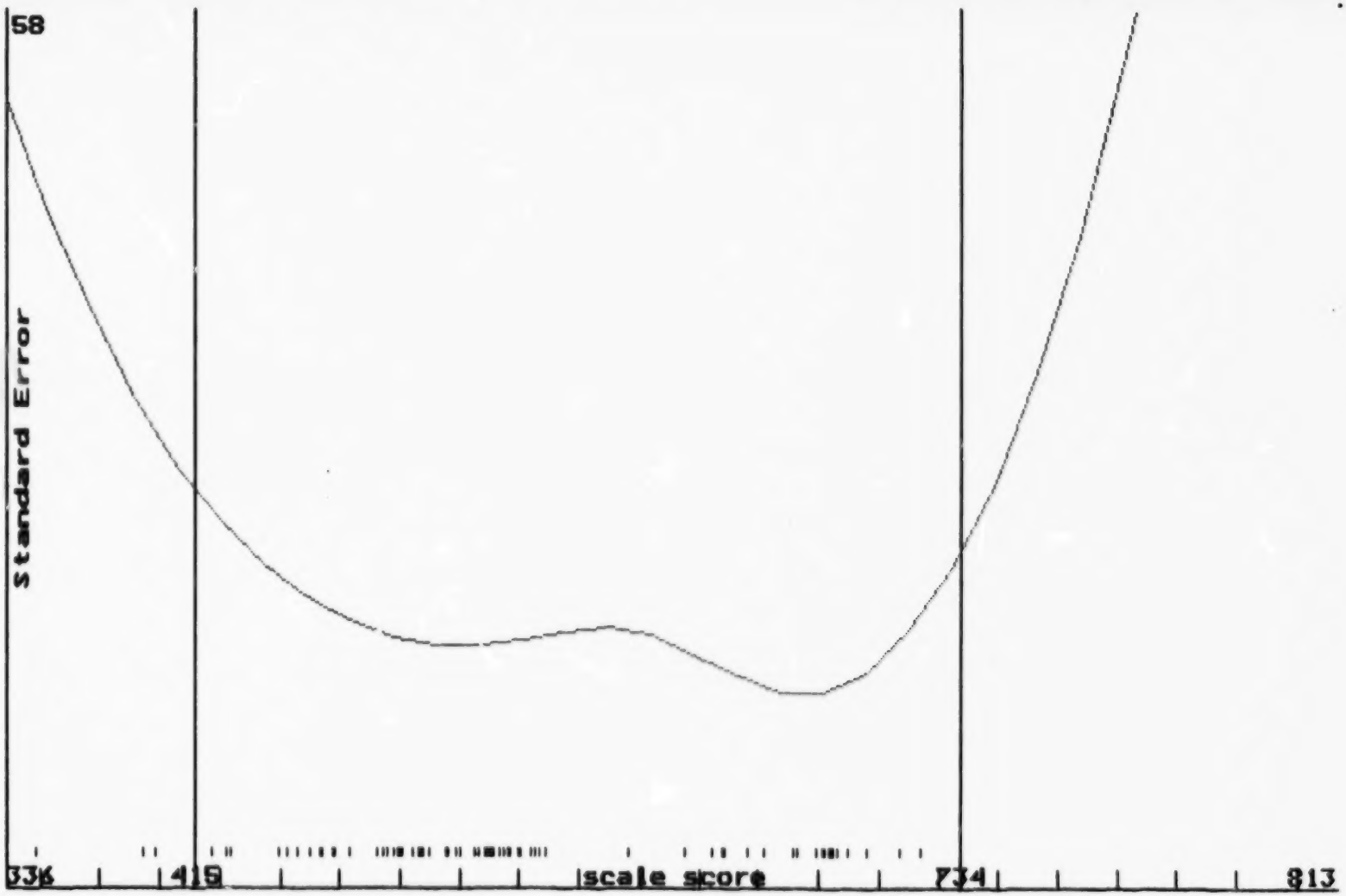

Figure 4. Standard Error Curve for an Item Pool with Insufficient Items in the Middle of the Scale

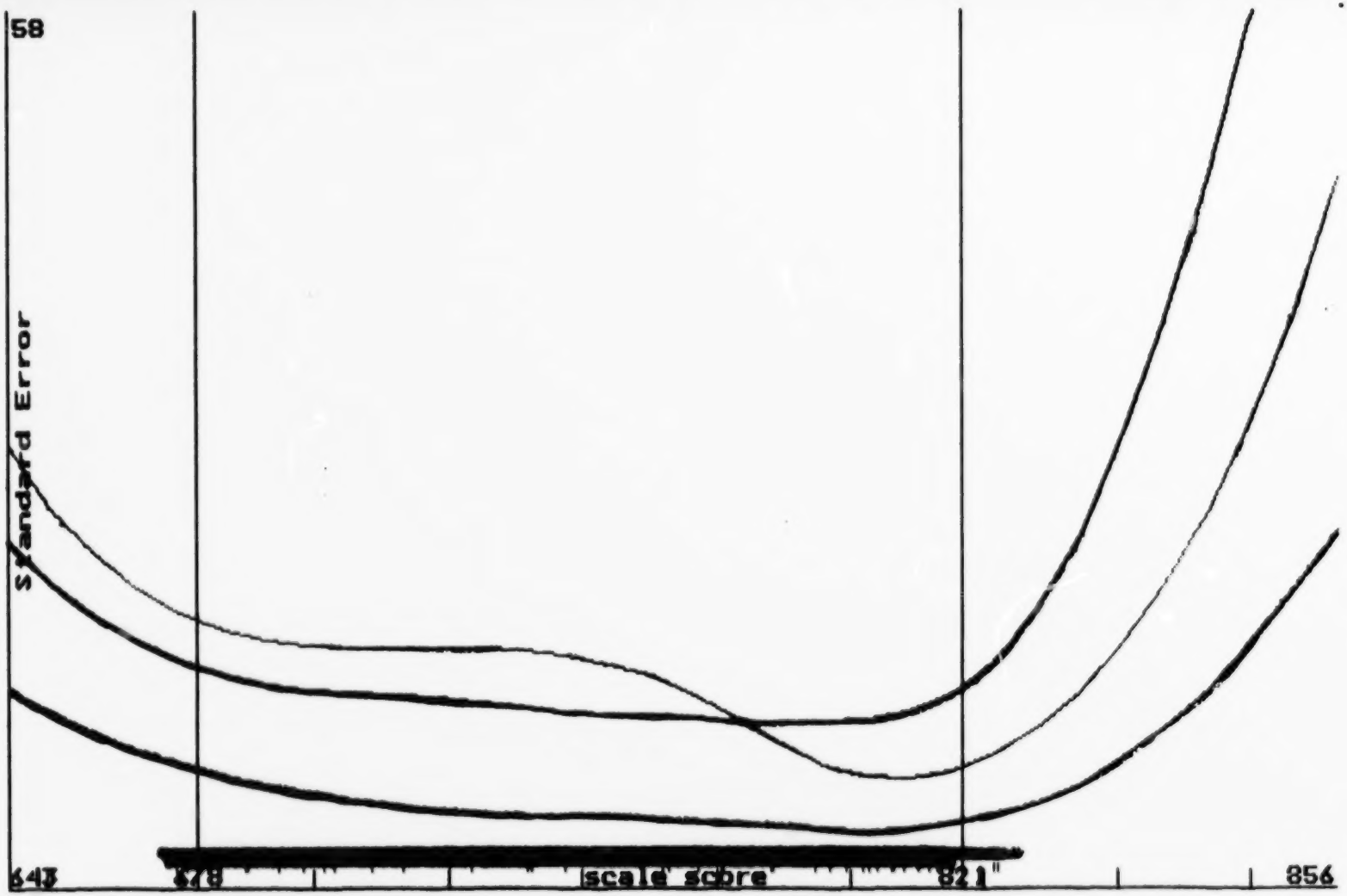

Figure 5. Standard Error Curves for a Pair of Automatically Selected Tests

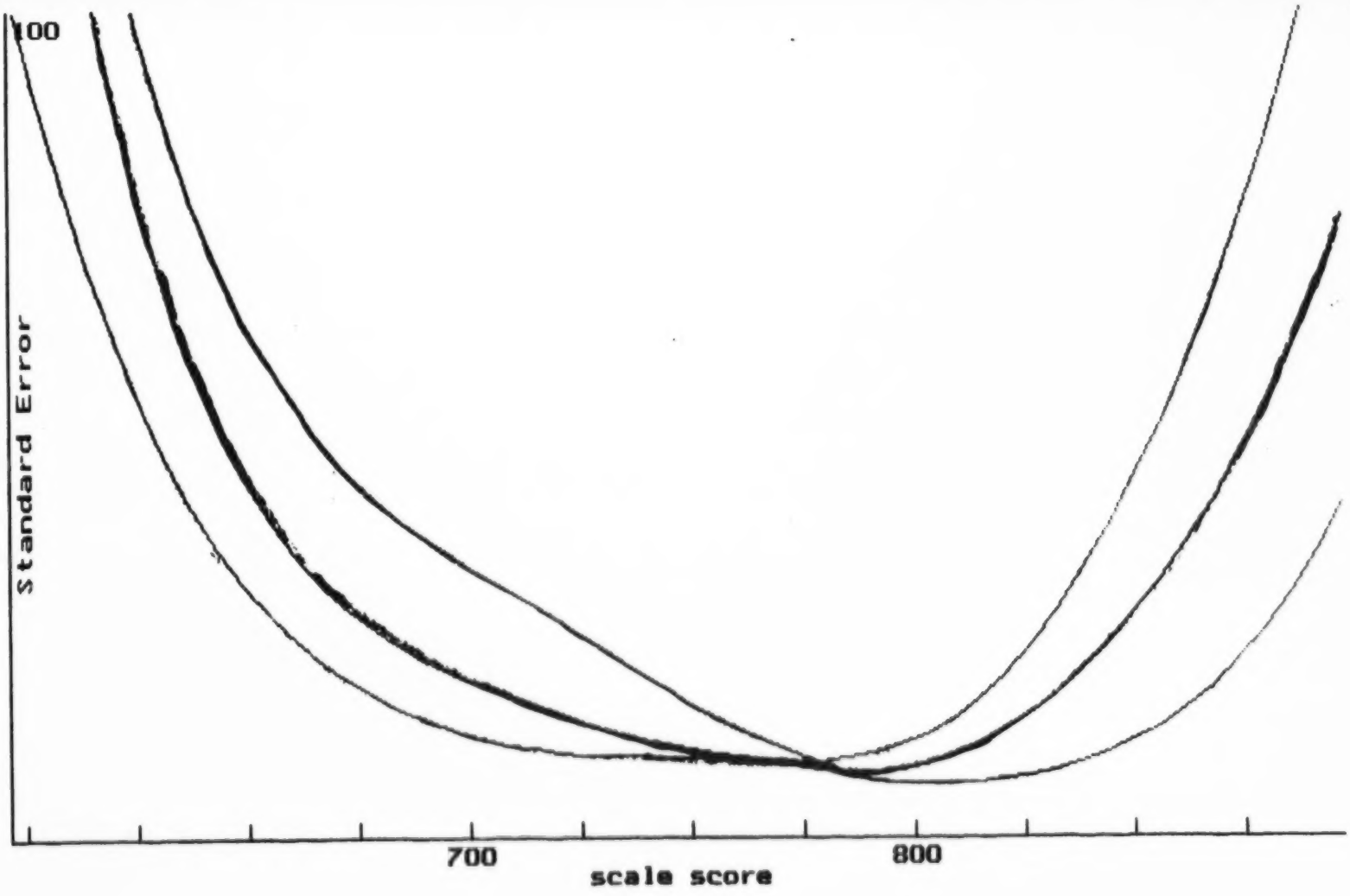

Figure 6. Standard Error Curves for Three Successive Levels of a Test